Trait evaluations of faces and voices: Comparing within- and between-person variability

Nadine Lavan[1,2], Mila Mileva[3,4*], A. Mike Burton[3], Andrew W. Young[3], & Carolyn McGettigan[1]*

** These authors contributed equally*

[1] Department of Speech, Hearing and Phonetic Sciences, University College London
[2] Department of Psychology, School of Biological and Chemical Sciences
Queen Mary University of London
[3] Department of Psychology, University of York
[4] School of Psychology, Faculty of Health: Medicine, Dentistry and Human Sciences, University of Plymouth

Correspondence to:
Nadine Lavan, Department of Psychology, School of Biological and Chemical Sciences
Queen Mary University of London, Mile End Road, London E1 4NS, United Kingdom.
E-mail: n.lavan@qmul.ac.uk
or
Mila Mileva, School of Psychology, Faculty of Health: Medicine, Dentistry and Human Sciences, University of Plymouth, Drake Circus, Plymouth PL4 8AA, United Kingdom. E-mail: mila.mileva@plymouth.ac.uk

Word Count: 11349

Funding: This work was supported by a Research Leadership Award from the Leverhulme Trust (RL-2016-013) awarded to Carolyn McGettigan and a Sir Henry Wellcome Fellowship (220448/Z/20/Z) awarded to Nadine Lavan

Author note: This manuscript was added as a preprint to PsyArXiv: 10.31234/osf.io/3rjc4
The data underlying the reported analyses have been uploaded to the OSF: https://osf.io/reb5j/

Abstract
Human faces and voices are rich sources of information that can vary in many different ways. Most of the literature on face/voice perception has focussed on understanding how people look and sound different to each other (between-person variability). However, recent studies highlight the ways in which the same person can look and sound different on different occasions (within-person variability). Across three experiments, we examined how within- and between-person variability relate to one another for social trait impressions by collecting trait ratings attributed to multiple face images and voice recordings of the same people. We find that within-person variability in social trait evaluations is at least as great as between-person variability. Using different stimulus sets across experiments, trait impressions of voices are consistently more variable within people than between people – a pattern that is only evident occasionally when judging faces. Our findings highlight the importance of understanding within-person variability, showing how judgements of the same person can vary widely on different encounters and quantify how this pattern differs for voice and face perception. The work consequently has implications for theoretical models proposing that voices can be considered 'auditory faces' and imposes limitations to the 'kernel of truth' hypothesis of trait evaluations.

Keywords: Trait perception, faces, voices, within-person variability

Introduction

We cannot help but form impressions of unfamiliar people's traits, that is, do we believe a person to be, for example, trustworthy, dominant or attractive. Depending on the specific circumstances, these impressions can be based on the way someone looks (e.g., when meeting them in person or seeing their photo on social media) or the way someone sounds (e.g., when speaking with them on the phone; McAleer, Todorov, & Belin, 2014; Todorov, Olivola, Dotsch & Mende-Siedlecki, 2015). Such snap judgements can often be made within milliseconds, but are nonetheless known to affect our behaviours, attitudes and decisions (Bar, Neta, & Linz, 2006; McAleer et al., 2014; Willis & Todorov, 2006). Most importantly, this is true in a range of situations, even when we have access to additional and more relevant information – from choosing who to vote for in the upcoming elections, to deciding the length and severity of court sentences, or the Airbnb host we decide to stay with (Chen, Halberstam, & Yu, 2016; Ert, Fleischer, & Magen, 2016; Klofstad, 2016; Mileva et al., 2020; Sussman, Petkova, & Todorov, 2013; Tigue, Borka, O'Connor, Schandl & Feinberg, 2012; Wilson & Rule, 2015). Despite their impact on many aspects of our lives, such social trait evaluations are unlikely to be firmly grounded in truth (Todorov et al., 2015; but see the 'kernel of truth' hypothesis, Berry, 1991), and the evidence for their accuracy is limited (Klofstad & Anderson, 2018; Todorov, 2017).

Intriguingly, however, trait impressions have been shown to be shared – that is, different people will usually agree, to some extent, on whether someone looks or sounds relatively trustworthy or not (Kramer, Mileva, & Ritchie, 2018; McAleer et al., 2014; Zebrowitz & Montepare, 2008). This indicates that there is some visual information in the human face and acoustic information in the human voice that we interpret in consistent ways to form trait evaluations. Moreover, social trait evaluations from both faces and voices have been shown to follow the same general structure, with trustworthiness and dominance being two principal dimensions underlying evaluation (McAleer et al., 2014; Oosterhof & Todorov, 2008). Some studies using natural face images and meaningful voice utterances (specifically "Hello") have also highlighted attractiveness as a potential third dimension underpinning social evaluations (McAleer et al., 2014; Sutherland et al., 2013).

Both faces and voices offer rich sources of information that can vary along multiple dimensions. These include cues related to invariant properties such as identity, age or sex, and transient cues such as emotional expressions, speaking style, eye gaze or head orientation, as well as external world cues such as lighting, background noise, distance from the recording device or its quality. Until recently, transient intrinsic and external-world cues were largely disregarded or controlled away in studies of person perception from faces and voices. However, recent work has highlighted the importance of these sources of variability in the context of both social evaluation and identity perception (Burton, 2013; Jenkins et al., 2011; Lavan et al., 2019; Todorov & Porter, 2014). This research highlights the valuable insights that can come from moving away from the standardised and strictly controlled face images or voice recordings predominantly used in the existing literature, and instead moving towards representing identities in a more naturalistic way. This enables better sampling of person variability and thereby better understanding of how faces and voices are processed in our everyday lives.

Accounting in empirical and theoretical work for the observation that the face and voice of the same person can vary from moment to moment is therefore essential to understanding person perception, as within-person variability has important implications for theories of social perception and identity recognition (Young, 2018; Young, Frühholz, & Schweinberger, 2020). Recent studies have already demonstrated that the variability in social trait evaluations attributed to different face photographs of the same person is indeed substantial, to the extent that it matches or sometimes even exceeds the variability in social ratings attributed to images of different people (Jenkins et al., 2011; Todorov &

Porter, 2014). This pattern of results has been reported for many different social traits, including attractiveness, trustworthiness, dominance, competence, creativity, and others, in both natural and more controlled face stimuli sets (Mileva et al., 2019; Sutherland, Young, & Rhodes, 2017). Estimating the relative proportion of within- and between-person variability in social ratings has important implications for the existing literature, which often takes a single (usually highly standardised) image as a veridical representation of a person. This leads to the implicit assumption that one identity can only be associated with one rating (e.g., of trustworthiness or dominance) and therefore any differences in social ratings are interpreted as cues to differences in identity. Moreover, collecting ratings attributed to images of the same person could introduce a novel approach to resolving the still ongoing 'kernel of truth' debate (Berry, 1991; Todorov et al., 2015). That is, if ratings of different images of the same person vary substantially, then these impressions would be limited in their ability to capture real and stable personality characteristics.

For voices, we know little about how within- and between-person variability relate to each other in trait evaluations, and the few studies available compare these two sources of variability in the acoustic properties of voices rather than their social evaluations. Kreiman et al. (2015) measured a number of acoustic properties, such as vocal pitch (F0) or harmonics-to-noise ratio, for nine different voice recordings of each of five different identities. This allowed them to explore the variation in these acoustic measures within the nine recordings of each person (within-person variability) as well as the variation across all five speakers (between-person variability). Their analysis showed more between- than within-person variability in the acoustics of the voices, although there were still substantial differences across recordings of the same person and even some cases where within-person variability exceeded between-person variability (see also Atkinson, 1976). However, the materials analysed were recordings of the sustained /a/ vowel, which does not adequately represent the within-person information usually apparent in daily life. In the current study, we aim to further the understanding of person perception by extending findings from the face perception literature and providing the first estimate of within- and between-person variability in social evaluations attributed to voice recordings.

The richness of the information human observers can derive from faces and voices has important implications. In particular, impressions of social traits and recognition of someone's identity have different underlying functional demands (Young et al., 2020); whilst perceiving within-person variability is critical to interpreting faces and voices to adequately form different impressions, it is something that needs to be discounted to recognise a person's identity. There is therefore a stark contrast between the role of within-person variability in identity perception and social evaluation. For identity perception, it has been observed that discriminating unfamiliar identities from either their face or voice can be surprisingly error-prone and this has been mostly attributed to our difficulties in processing between-person variability, or "telling people apart" (Hancock, Bruce, & Burton, 2000; Kreiman & Sidtis, 2011; Lavan, Scott, & McGettigan, 2016; Read & Craik, 1995; Young & Burton, 2018). However, Jenkins et al. (2011) showed that the difficulties we experience also affect "telling people together" – when presented with natural within-person variability, people often misinterpret two images of the same (unfamiliar) face as depictions of two different people. Similar findings have also been reported with sorting tasks using naturally-varying voice recordings (Lavan, Burston & Garrido, 2019; Stevenage, Symons, Fletcher & Coen, 2020), altogether highlighting how difficult it is for unfamiliar viewers and listeners to cope with the image and voice variability within a single person.

While large within-person variability hinders unfamiliar identity perception, these superficial differences in the way someone looks or sounds are directly relevant to the impressions we form. For example, a change in emotional expression or a change in vocal pitch can dramatically affect our perceptions of trustworthiness and dominance (Mileva et al., 2018; Ohala, 1982; Said, Sebe, &

Todorov, 2009) and both of these factors can vary across different encounters with the same person. These differences in how within-person variability may affect identity perception and social evaluation are particularly important in the context of reported similarities between face and voice perception.

Many aspects of face and voice processing are already incorporated in theoretical models (Belin et al., 2011; Bruce & Young, 1986; Yovel & Belin, 2013), including identity recognition, emotion perception, and speech perception. Social evaluations on the other hand are not as explicitly integrated in models, with theoretical work to date mainly focussing on evolutionary approaches. At the same time, theoretical models of voice perception have historically been heavily influenced by earlier face perception models and therefore highlight parallels between face and voice perception. This has led to the popular analogy of describing the voice as an 'auditory face'. Adopting this popular view, the effects of within-person variability on identity perception described above present yet another parallel between faces and voices – both viewers and listeners struggle to 'tell unfamiliar people together' (Jenkins et al., 2011; Lavan et al., 2019). This large within-person variability due to the superficial image and acoustic differences also affects first impressions. The 'auditory face' analogy would also predict that the formation of social evaluations from faces and voices should work in parallel.

In a series of three experiments, we set out to quantify the within- and between-person variability in social trait evaluation from faces and voices and examine how within- and between person variability relate to each other, thus empirically testing the 'kernel of truth' hypothesis. We furthermore directly compared social evaluations based on face and voice cues in the same study - a rarely adopted approach given the rather independent face and voice perception literatures. This has therefore allowed us to test the degree to which a voice may indeed be an 'auditory face' in this context.

In Experiment 1, we analysed data for social evaluations of trustworthiness, dominance, and attractiveness, which have been highlighted by previous research as the key dimensions for trait perception from faces and voices (Oosterhof & Todorov, 2008; Sutherland et al., 2013). In these data, we quantified the within- and between-person variability in perceived traits using a novel measure based on standard deviations calculated from evaluations. In Experiment 2, we sought to manipulate the amount of between-person variability in social judgements using stimulus selection, in order to test the generalisability of our findings. Finally, in Experiment 3, we extracted the visual and auditory information from multiple video recordings of the same person to directly compare the proportion of within- and between-person variability in trait evaluations for the same instances of faces and voices. In all three experiments, we sample naturally-occurring (ambient) face images and dynamic videos as well as voice recordings in a familiar and unfamiliar language, thus estimating the variability in faces and voices from a range of different stimuli and attempting to approximate the true variability we are presented with in daily life.

Experiment 1: Variability in trait ratings attributed to faces and voices

In our first experiment, we aimed to assess how within- and between-person variability in social evaluations relate to each other for faces and voices. Different groups of participants rated 20 face or voice stimuli of 20 unfamiliar identities (10 female) for trustworthiness, dominance, and attractiveness on a 9-point Likert scale. For the trait ratings of faces, we describe a re-analysis of existing data of social evaluations of faces reported in Mileva et al. (2019, Study 1). Details of the participants recruited and methods used are described again below. For trait ratings of voices, we collected new data with naturally-varying voice recordings, using a similar experimental design as used for the face ratings. Overall, we predicted that within-person variability would be higher or

similar compared to between-person variability, for both faces (see Mileva et al., 2019; Sutherland et al., 2017; Todorov & Porter, 2014) and voices (Atkinson, 1976; Kreiman et al., 2015).

Method

Participants

For social trait ratings from faces, 20 participants (mean age = 20.1 years, SD = 1.4 years) were recruited from the University of York. All participants had normal or corrected-to-normal vision and received payment or course credit for their participation. Racial identity was not controlled, although the majority of the students at the University of York are white. Sample size was based on previous studies collecting social ratings attributed to different images of the same identities (Todorov & Porter, 2014). Experimental procedures were approved by the Ethics Committee of the University of York Psychology Department and participants provided informed consent.

For social trait ratings from voices, 60 participants (32 female) aged between 18 and 35 years (mean = 27.3 years, *SD* = 5.7 years) were recruited for online testing via Prolific.co. All participants were native speakers of English. No participant had any self-reported hearing impairments. All participants responded correctly to over 80% of the catch trials (see *Procedure*; 59 participants = 100%, 1 participant = 80%). Ethical approval was given by the UCL Research Ethics Committee (Project ID number: SHaPS-2019-CM-030). The participants were randomly assigned to rate one of the three social traits (trustworthiness, dominance, and attractiveness), leading to a sample size of 20 participants per trait evaluation. Thus, social trait is a between-subjects factor for faces while it was a within-subjects factor for voices.

Materials

A total of 400 face images were used to collect social trait ratings attributed to faces. This included 20 different images of each of 20 unfamiliar white identities (10 female, age range = 22-54 years). These were minor celebrities from foreign countries, chosen to be unfamiliar to the UK participant pool. All images were collected via a Google Image Search by entering the name of the celebrity and selecting the first images that showed the whole face with no parts obscured by clothing or glasses. Images were naturally occurring (or ambient, Jenkins et al., 2011) and therefore included a large amount of variability due to lighting, pose, and emotional expression. Example images of the identities used throughout this experiment can be seen in Figure 1.

For social trait ratings from voices, we aimed to broadly match the properties of the stimulus set used for the trait ratings from faces. We therefore included 20 voice recordings from 20 different identities (10 female) from the LUCID corpus (Baker & Hazan, 2011). All speakers were monolingual speakers of Standard Southern British English, aged between 17-28 years. Voice recordings were single words that were extracted from connected speech produced in the context of a range of tasks and recording sessions. Specifically, 10 out of the 20 stimuli per identity were sampled from spontaneous speech elicited across a number of sessions and speaking environments through a “Spot the difference” task. The remaining 10 stimuli were sampled from fully or partially-scripted speech, elicited via sentence reading and picture naming tasks. Five of these scripted stimuli were produced in casual speech “as if talking to a friend”, the remaining five in clear speech, “as if talking to someone who is hearing-impaired” (Baker & Hazan, 2011). This stimulus selection process was aimed at including substantial within-person variability in the sampled voice recordings. The average duration of the voice recordings was 589ms (*SD* = 163ms) and the intensity of each stimulus was root-mean-square normalised. WAV files were then converted into MP3s for use in the online study.

Figure 1. Example images of the face identities used in Experiment 1, see also Mileva et al. (2019, Study 1). Restrictions prevent publication of the original images used in the experiment. Images included in the figure feature computer-generated images created using style-based GAN architecture (Karras et al., 2020) that do not represent existing identities but are comparable to the images used in the experiment. Each image shows a different identity. See image attributions in the Acknowledgements section.

Procedure

For social trait ratings of faces, participants rated all 400 images for three different social traits (trustworthiness, dominance, and attractiveness) on a 9-point Likert scale (1 = not at all trustworthy/dominant/attractive; 9 = extremely trustworthy/dominant/attractive) using a mouse-click. Each image was rated for a single social trait in a block of 400 trials, and the order of the blocks corresponding to each of the three traits was randomised. Thus, each participant completed 1200 trials. Image presentation order within each block was further randomised individually for each participant. Images were presented using MATLAB and the Psychophysics Toolbox (Brainard, 1997) and were displayed on an 18-inch LCD monitor. For each trial, an image was presented at the centre of the screen with a rating scale positioned underneath. The task was self-paced, with an inter-stimulus interval of 1 second, however, participants were encouraged to rely on their "gut instinct" (cf. Todorov, Mandisodza, Goren, & Hall, 2005) when making their judgements.

For social trait ratings from voices, our experiment was created and hosted on the Gorilla Experiment Builder (www.gorilla.sc, Anwyl-Irvine, Massonnié, Flitton, Kirkham & Evershed, 2018). After

providing informed consent, all listeners completed a headphones screening task where they were presented with three pure tones and asked to judge which one was the quietest (Woods, Siegel, Traer & McDermott, 2016). This ensured that participants were able to hear the voices adequately throughout the task. Following this screening, listeners were randomly assigned to rate all 400 recordings for one of the three social traits in the study. Thus each participant completed 400 trials. Listeners were first presented with a voice recording and after hearing the recording in full, a rating scale appeared on the screen. This rating scale ranged from 1-9 ("How trustworthy/dominant/attractive does this person sound?" 1 – not trustworthy/dominant/attractive at all; 9 – very trustworthy/dominant/attractive). The order of the stimuli was fully randomised for each participant. The task was self-paced and participants took around 20 minutes to complete it. To ensure that participants were continuously paying attention to the task and listening to the stimuli, 12 catch trials were included at random intervals, for which a voice instructed listeners to press a certain number on the rating scale (e.g. "Please click on 1.").

Quantifying within- and between-person variability

We quantified the variability in our data sets in a similar way to the procedure outlined by Todorov and Porter (2014, also used in Sutherland et al. 2017). For each participant, we calculated a single estimate of the within- and the between-person variability in their social ratings based on standard deviations. We calculated these variability estimates separately for the 10 male and 10 female identities. Since previous studies suggest some sex-specific effects in social trait ratings (see Mileva et al., 2019; Sutherland et al., 2015; Todorov & Porter, 2014) computing variability estimates by sex is desirable to avoid inflating the between-person variability. Thus, our variability estimates use the standard deviation as opposed to the related measure of variance which is used by Todorov and Porter (2014) and we calculate variability estimates per participant as opposed to per social trait and sex.

Specifically, we quantified the within-person variability in the following way for all experiments reported in this paper:

1. For each participant, and each trait, we computed the standard deviation in ratings for the different stimuli that represented each identity, resulting in identity-specific estimates of the within-person variability.
2. We then averaged these identity-specific estimates of the within-person variability across the (i) 10 female and (ii) 10 male identities, to obtain a single within-person variability estimate per sex.

We quantified the between-person variability in the following way for all experiments reported in this paper:

1. For each participant, and each trait, we averaged the ratings of the different stimuli per identity to obtain an identity-specific trait rating.
2. We then computed the standard deviation between these identity-specific average trait ratings for each sex, resulting in a single between-person variability estimate for male and female identities.

Thus, we computed an estimate of the within-person variability and an estimate of the between-person variability for male and female identities, separately for each participant.

For this measure, the lowest possible estimate for either type of variability is 0, where all 10 data points are associated with the same value (e.g. all 10 images/recordings of an identity were rated as 4 for one of the social traits). The highest possible estimate is 4.22, where half the data points fall on one extreme of the scale and half on the other (e.g. 5 images/recordings of an identity were rated as 1 and the other 5 images/recordings were rated as 9).

For Experiment 1 only, we implemented further steps to match the number of data points entered into the estimation of the within- and between-person variability. Since the variability estimates were computed by sex, there was a potential imbalance in the number of data points included in the computations of the within-person variability (20 stimuli per identity) and the between-person variability (10 identities per sex). To address this problem and match the number of data points, we iteratively (1000 permutations) drew random samples of 10 stimuli per identity to compute the standard deviations per identity and compute the within-person variability. For the between-person variability, we in turn computed the identity-specific average trait ratings based on 1000 random samples of 10 stimuli per identity to form the basis of the between-person variability estimates. The final estimates presented here are the average within- and between-person variability estimates across these 1000 permutations respectively.

Results and Discussion

Inter-rater reliability

For faces, ratings for all three social traits showed good inter-rater agreement with Cronbach's $\alpha > .88$. We calculated Cronbach's alpha because it is the most widely reported statistic in the literature. However, because using Cronbach's alpha as a measure of inter-rater agreement might be problematic (Cortina, 1993), we also calculated the intraclass correlation coefficient (ICC) as a more appropriate measure of agreement, separately for ratings of each trait. We used the Two-Way Random model as all participants rated all face images used in the experiment and report the values for absolute agreement. These analyses showed significant rater agreement for ratings of trustworthiness (ICC = .698, 95% CI [.65, .74], $p < .001$), dominance (ICC = .727, 95% CI [.68, .77], $p < .001$), and attractiveness (ICC = .945, 95% CI [.93, .95], $p < .001$).

For voices, ratings across the three social traits showed good inter-rater agreement with Cronbach's $\alpha > .70$, although we note these were somewhat lower compared to face ratings. We also calculated the ICC using a Two-Way Random model and absolute agreement. This analysis showed significant rater agreement for ratings of trustworthiness (ICC = .649, 95% CI [.59, .70], $p < .001$), dominance (ICC = .695, 95% CI [.65, .74], $p < .001$) and attractiveness (ICC = .783, 95% CI [.74, .82], $p < .001$).

Comparing the within- and between-person variability in social trait ratings

To assess how within- and between-person variability relate to one another, we ran linear mixed models (LMMs) using the *lme4* package (Bates, Maächler, Bolker, & Walker, 2015) in the R environment. All categorical predictors were dummy-coded before being entered into the LMMs. The LMMs had variability type (within vs between) and social trait (trustworthiness, dominance, and attractiveness) and sex as fixed factors and participant was entered as a random effect. No random slopes were added as the inclusion of such effects led to models not converging. All interactions between variability type and social trait were modelled. Sex was not a factor of interest and was not modelled in the interactions.

*lmer(variability estimate ~ variability type * social trait + sex + (1|participant)*

We ran separate models for faces and voices as social trait was a between-subjects factor for faces while it was a within-subjects factor for voices. We note that in the following experiments, different assignment of the within- vs between-subject factors enabled us to include the data for faces and voices in the same model. Significance of the main effects and interactions was established via log-likelihood tests by dropping effects of interest from the appropriate model. For example, to test for the significance of the two-way interactions we dropped this interaction from the model, only retaining the two main effects. For additional plots of mean ratings per image and identity, please refer to Supplementary Figure 1.

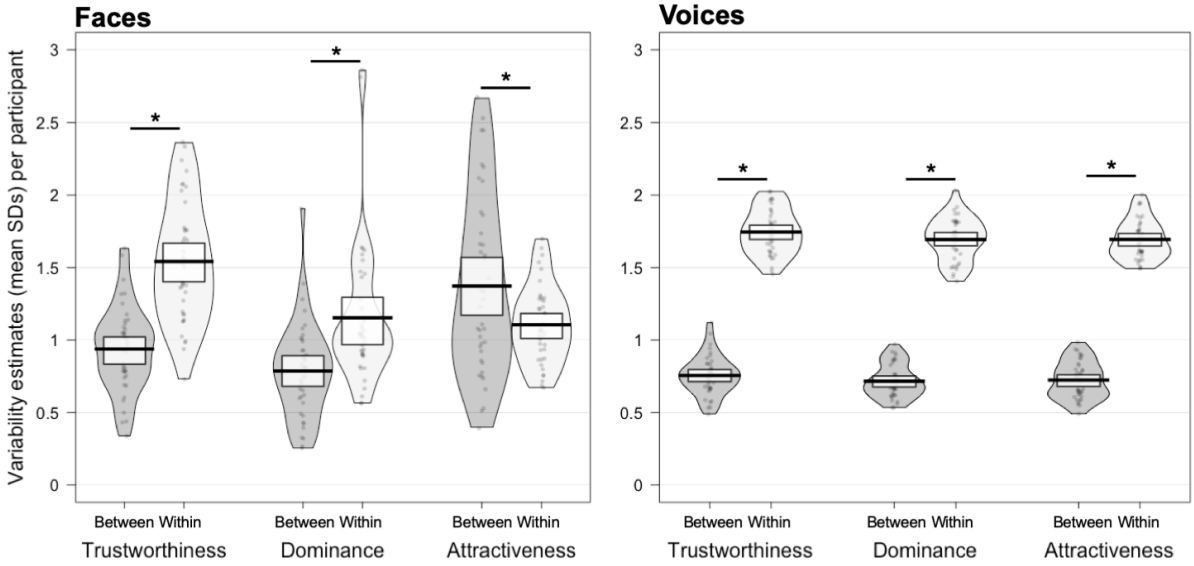


Figure 2. Within- and between-person variability in ratings of trustworthiness, dominance, and attractiveness attributed to faces and voices in Experiment 1. Boxes indicate the 95% confidence intervals. * indicates $p < .05$. Please note that the data underpinning the plot for faces were collected as part of another study (Mileva et al., 2019) and was re-analysed for this paper.

For faces, one datapoint measuring within-person variability in dominance ratings was identified as an outlier (being > 3 SDs above the mean). The following analysis was conducted with this datapoint included, as removing it did not affect the results. There was a significant interaction between variability type and social trait ($\chi^2[2] = 46.67$, $p < .001$), such that the relationship between within- and between-person variability differed by social trait. The main effect of trait was significant ($\chi^2[2] = 19.19$, $p < .001$), as was the main effect of type of variability ($\chi^2[1] = 16.54$, $p < .001$). In the presence of an interaction, we, however, note that these main effects are of limited interpretability. We used *emmeans* to run pairwise post-hoc tests on these models. Degrees of freedom were calculated using the Kenward-Roger Approximation. These post-hoc tests showed that within-person variability exceeded between-person variability for trustworthiness and dominance (*t*s[214] > 4.11, *p*s < .001) but that between-person variability exceeded within-person variability for attractiveness ($t[214] = 2.98$, $p = .003$).

For voices, there was no significant interaction between variability type and social trait ($\chi^2[2] = .32$, $p = .854$) and no main effect of social trait ($\chi^2[2] = 2.97$, $p = .226$). There was however a main effect of variability type ($\chi^2[1] = 661.67$, $p < .001$), such that the degree of within- and between-person variability differed in similar ways for all traits. Post-hoc tests conducted in *emmeans* showed that within-person variability exceeded between-person variability for all three social traits (*t*s[176] > 37.79, *p*s < .001). Please see the supplementary materials for additional analyses showing that there was no clear effect of linguistic content on trait ratings (Supplementary Analysis 1).

We therefore find that within-person variability in social trait evaluations for faces and voices is substantial. For faces, the within-person variability exceeded the between-person variability for trustworthiness and dominance, largely replicating the pattern of previous findings in the literature through quantifying the within- and between-person variability via standard deviations per participant (Sutherland et al., 2017; Todorov & Porter, 2014). For voices, the within-person variability in social trait evaluations was also sizeable, echoing our findings for variability in trait evaluations from faces. However, in contrast to our findings for faces, within-person variability always exceeded between-person variability for impressions from voices. From this data set, this effect appears to be due both to within-person variability being more pronounced for voices than for faces, while at the same time

between-person variability is reduced (see Figure 2). Unexpectedly and in contrast to the data from faces, all social traits behave in similar ways for voices. Nevertheless, we note that the face and voice stimuli were sampled from quite different source materials and consequently with different selection strategies (see the Materials for Experiment 1). All of these factors could have influenced the amount of within- and between-person variability in the ratings attributed to the face and voice identities. We return to this point in Experiment 3 where we specifically control for such potential discrepancies in the face and voice stimuli.

Overall, we provide further empirical evidence that the same person can create very different impressions from how they look or sound; e.g. very trustworthy in one instance but very untrustworthy in another. This is demonstrated by our findings that the variability in social trait evaluations for two images or recordings of the same person can be similar to (or larger than) the differences in evaluations of images or recordings of two different people.

Experiment 2: Manipulating within- and between-person variability of faces and voices

In Experiment 2, we tested whether the absolute estimates of the within- and between-person variability are specific to the chosen stimulus set, or generalisable and thus representative of population-wide estimates. For example, the voices included in Experiment 1 were all produced by speakers with a Standard Southern British English accent. This makes this particular set of voices fairly homogeneous, which may be reflected in the relatively low between-person variability. Conversely, we hypothesised that the properties of the specific set of faces used in Experiment 1 might also have contributed to the larger estimates of between-person variability: For example, the set of faces included identities of different ages, hair and eye colours, and complexions, all of which could affect the perceived variability in social evaluations attributed to images of different identities.

We therefore created new stimulus sets for Experiment 2 that aimed to reduce between-person differences in faces and increase between-person differences in voices. We selected a set of faces, in which the individual identities were more homogenous in visual presentation (white females with long blonde hair and light eyes) to examine whether the degree of between-person variability for faces in relation to the within-person variability would be reduced.

For voices, we selected a set of identities with less homogeneous-sounding voices (speakers with a variety of pronounced regional UK accents) to examine whether the degree of between-person variability for the voices in relation to the within-person variability would increase, thus more closely resembling the pattern of results observed for faces in Experiment 1. We specifically chose to include identities with different regional UK accents as there is evidence that such accents are evaluated differently along many social evaluation dimensions. For example, speakers with a Birmingham accent have been perceived as less intelligent than speakers with the standard RP British accent (Giles, Wilson, & Conway, 1981), whereas speakers with a South Welsh accent have been perceived as kinder than speakers with the standard RP British accent (Giles, Baker, & Fielding, 1975, see also Fuertes et al., 2012 for a review). If the findings in Experiment 1 are indeed representative of face and voice processing in general, then we should see the same broad pattern in how within- and between-person variability in trait ratings relate to each other in the following experiment.

Method

Participants

For social trait ratings of faces, 60 participants (28 female) aged between 18 and 35 years (mean = 27.9 years, SD = 5.9 years) were recruited for online testing via Prolific.co. Racial identity was not controlled, although the majority of the participants on Prolific.co are white. All participants responded correctly to over 80% of the catch trials (57 participants = 100%, 3 participants = 90%).

For social trait ratings of voices, 62 participants aged between 18 and 35 years were recruited for online testing via Prolific.co. All participants were native speakers of English. No participant had any self-reported hearing impairments. From this sample, two participants were excluded because our catch trials indicated that they were not paying sufficient attention to the task (errors on > 20% of the catch trials). For the remaining sample, 58 participants responded with 100% accuracy on the catch trials and 2 participants responded correctly for 80%. This resulted in a final participant sample of 60 listeners (mean = 27.2 years, *SD* = 5.7 years; 32 female).

Participants were randomly assigned to rate one of the three traits (trustworthiness, dominance, and attractiveness), leading to a sample size of 20 participants per trait rating. Ethical approval was given by the UCL Research Ethics Committee (Project ID number: SHaPS-2019-CM-030).

Materials

Since the effects of sex of the identity is not the main focus of the current experiment, we only included female identities to streamline the data collection process.

For the social trait ratings of faces, we selected a total of 100 images (10 images of each of 10 unfamiliar identities, age range = 30-54). In order to reduce the amount of between-person variability in the ratings attributed to the face images, all identities were specifically selected to follow a common physical description – white female with long blonde hair and light-coloured eyes (1 exception with brown eyes. Note that the reported results below do not change when excluding the brown eyed identity). As in Experiment 1, all images were downloaded with a Google Image Search by entering the name of the identity (a foreign celebrity unknown in the UK) and selecting the images that fit our pre-established criteria as well images where no parts of the face were obscured by clothing or accessories. All images were again naturally occurring (or ambient, Jenkins et al., 2011) and varied in lighting, pose, and emotional expression among other properties. Figure 3 shows example images of some of the identities used in Experiment 2.

Figure 3. Example images of the face identities used in Experiment 2. Each image shows a different identity. Restrictions prevent publication of the original images used in the experiment. Images included in the figure feature computer-generated images (Karras et al., 2020) that are comparable to the images used in the experiment.

For the social trait ratings from voices, we selected 10 female speakers from the 9 distinctive regional UK accents included in the IViE corpus (Nolan & Post, 2014). All speakers were 16 years old. The accents included in the study were Belfast (two speakers), Bradford, London, Cambridge, Leeds, Cardiff, Dublin, Liverpool and Newcastle. Note that we did not formally check the familiarity of our participants with these accents but assume that the distinctive phonetic, prosodic, and linguistic (for spontaneous speech) properties of this set of dialects should be readily perceived as more variable than the single standard accent used in Experiment 1. Voice clips were selected evenly from a range

of tasks to ensure that sufficient within-person variability was sampled (see Experiment 1). These tasks included sentence reading, passage reading, spontaneous retelling of a previously read passage, spontaneous conversational speech, and a conversational spot the difference task. All stimuli were full meaningful utterances, with an average duration of 1.43 seconds (*SD* = .33 seconds). All stimuli were root-mean-square normalised for intensity.

Procedure

This experiment was created and hosted on the Gorilla Experiment Builder (www.gorilla.sc, Anwyl-Irvine et al., 2018). Participants were randomly assigned to rate all 100 stimuli for one of the three traits in this study. Thus, each participant completed 100 trials. Images were presented in the centre of the screen, with a rating scale underneath the image. This rating scale ranged from 1-9 ("How trustworthy/dominant/attractive does this person look?" 1 – not trustworthy/dominant/attractive at all; 9 – very trustworthy/dominant/attractive). The order of the stimuli was randomised for each participant. The task was self-paced and participants took around 8 minutes to complete it. To ensure that participants were continuously paying attention to the task, catch trials were included, for which a number between 1 and 9 was presented on the screen instead of a face image. For these trials, participants were instructed to select this number on the rating scale. All participants responded correctly on more than 80% of these catch trials and thus no participants were excluded based on the catch trials.

For voices, participants also rated all 100 stimuli for one of the three social traits. The procedure was otherwise identical to the one reported for voices for Experiment 1.

Results and Discussion

Inter-rater reliability

For faces, ratings for all three social traits showed good rater agreement with Cronbach's $\alpha > .94$. We also calculated the ICC using a Two-Way Random model and absolute agreement. This analysis showed significant rater agreement for ratings of, trustworthiness (ICC = .828, 95% CI [.77, .88], $p < .001$), dominance (ICC = .830, 95% CI [.77, .88], $p < .001$), and attractiveness (ICC = .839, 95% CI [.78, .89], $p < .001$).

For voices, ratings for all three traits showed good inter-rater agreement with Cronbach's $\alpha > .96$. As in the previous experiments, we also calculated the ICC using a Two-Way Random model and absolute agreement. This analysis showed significant rater agreement for ratings of trustworthiness (ICC = .587, 95% CI [.47, .69], $p < .001$), dominance (ICC = .832, 95% CI [.78, .88], $p < .001$) and attractiveness (ICC = .648, 95% CI [.54, .74], $p < .001$).

Comparing the within- and between-person variability in social trait ratings

Variability estimates were obtained in a similar way as described in the previous experiment. For the current experiment, since each identity was represented by 10 stimuli, no permutations were necessary (cf Experiment 1). Therefore, the within-person variability was estimated by first calculating the standard deviation of the raw ratings of the stimuli per identity and then averaging these identity-specific standard deviations to obtain a single within-person variability estimate per participant. The between-person variability was computed by averaging the raw ratings per identity, thus computing a mean trait rating representation of the identity as opposed to the different stimuli. From these identity-specific averages, we then computed to standard deviation, to again obtain a single between-person variability estimate from each participant. The mean variability estimates are shown in Figure 4 for all three social traits.

To assess how within- and between-person variability relate to one another, we ran linear mixed models (LMMs) using the *lme4* package (Bates et al., 2015) in the R environment. The LMMs had variability type (within vs between), social trait (trustworthiness, dominance, and attractiveness) and also modality (faces vs voices) as fixed factors and participant as a random effect. All interactions between fixed factors were modelled.

*lmer(variability estimate ~ variability type * social trait * modality + (1|participant)*

Significance of the main effects and interactions was established via log-likelihood tests in the same manner as for the models reported in Experiment 1. For plots of mean ratings per image and identity, please refer to Supplementary Figure 2.

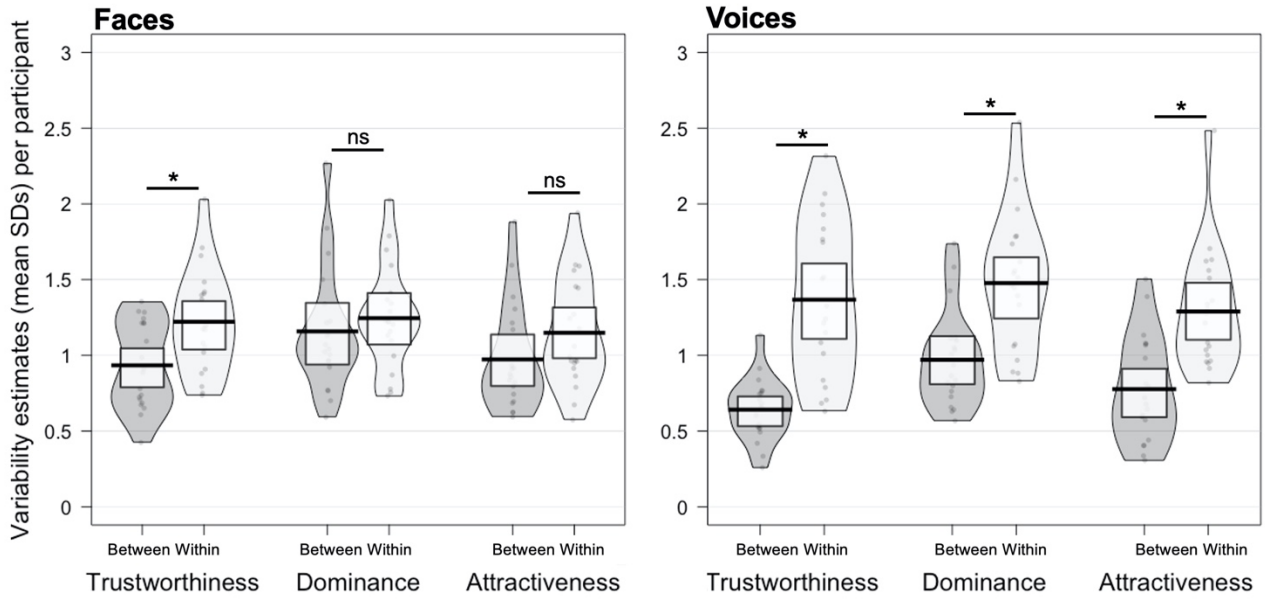


Figure 4. Within- and between-person variability in ratings of trustworthiness, dominance, and attractiveness attributed to faces and voices in Experiment 2. Boxes indicate the 95% confidence intervals. * indicates $p < .05$.

The three-way interaction between social trait, modality, and variability type was not significant (χ^2[2] = .025, p = .880). There were similarly no two-way interactions between variability type and social trait (χ^2[2] = 4.21, p = .122) and modality and social trait (χ^2[2] = .80, p = .671). The interaction between variability type and modality was, however, significant (χ^2[1] = 19.75, $p < .001$), indicating that the degree of difference between within- and between-person variability estimates depends on the stimulus modality. There was finally a significant main effect of social trait (χ^2[2] = 8.36.75, p = .015). Planned post-hoc tests conducted in *emmeans* showed that for faces, within-person variability exceeded between-person variability for trustworthiness (t[173] = 3.47, $p < .001$) but was on par with between-person variability for dominance (t[173] = 1.10, p = .27) and attractiveness (t[173] = 1.63, p =.10). For voices, within-person variability exceeded between-person variability for all three traits (ts[173] > 5.58, ps < .001). The difference between within- and between-person variability is therefore overall larger for voices, driving the interaction.

As in Experiment 1, within-person variability was either on par with or exceeded the between-person variability for both faces and voices. For faces, contrary to our predictions, however, the between-person variability was not systematically reduced in this visually more homogeneous set of face stimuli (see Figures 1 and 3). If anything, the between-person variability was slightly increased in Experiment 2 and, crucially, the pattern of results (within-person variability being on par with the

between-person variability for dominance and attractiveness and exceeding between-person variability for trustworthiness) is similar to what we observed in Experiment 1. The specific sets of ambient images and face identities used thus seem to have only a small effect on the overall variability in social trait evaluations.

For voices, within-person variability systematically exceeded between-person variability for all three social trait evaluations of voices. In line with predictions, the between-person variability appears somewhat increased compared to Experiment 1, although, against predictions, the within-person variability appears to be reduced. Introducing regional accents in the stimuli to increase the between person-variability in trait evaluations therefore did not have a dramatic effect on the overall variability in listeners' judgements. In fact, the pattern of results is again similar to Experiment 1.

The stimulus manipulations introduced in Experiments 2 thus overall only resulted in small changes in the amount of variability in social trait evaluations. Faces that look broadly similar (blonde women with long hair) are not necessarily perceived to be more similar in social traits than faces that look less similar to one another. Voices with the same (standard) accent are not perceived as dramatically more similar to one another in terms of their social traits than voices with very different, regional accents. Crucially, the relationship of within- and between-person variability for the three trait evaluations remained the same across experiments for faces and voices respectively. Thus, the similarity of results across Experiments 1 and 2 speak to a certain degree of generalisability of the relative weighting of within- vs between-person variability in faces and voices.

Experiment 3: Variability in trait ratings attributed to the face and voice of the same identities

In Experiments 1 and 2, the relationship of within- and between-person variability across social trait evaluations appears to differ systematically for faces and voices, such that within-person variability more reliably exceeds between-person variability for impressions from voices.

However, these apparent differences across modalities may have been introduced by some of the design choices in these previous experiments: 1) Experiments 1 and 2 used static images of faces and dynamic recordings of voices, 2) The images and voice recordings were sampled from different sets of identities, 3) Face images represented a distinct encounter with each identity (i.e., each picture was taken on a different day or at a different event), whereas all voice samples were recorded on the same day and session, 4) Linguistic information was accessible/present in the voice stimuli but not in the face stimuli, 5) Face and voice rating experiments were run independently of one another, with different participants completing each task. For all of these reasons, directly comparing the results for faces and voices might be of limited validity.

In Experiment 3, we aimed to eliminate these potential confounds to directly investigate modality-specific differences in the variability of trait ratings. We therefore selected a new set of stimuli with the following properties: 1) We included dynamic videos of faces to match the dynamic nature of the voice stimuli, 2) We used the same identities for faces and voices, 3) We extracted the video and audio content from the same original videos, such that each video and audio clip then represented a (common) distinct encounter with the identity, 4) We used voice recordings of Bulgarian celebrities speaking in a language unfamiliar to our participants to avoid any potential interference from linguistic content, and 5) We implemented a within-subjects design, where the same participants rated both faces and voices for the same social traits.

Based on the findings from Experiments 1 and 2, we predicted that we should observe similar trends, where within-person variability in trait ratings is equivalent to, or exceeds, between-person variability.

However, we expect that the tendency for within-person variability to exceed between-person variability would be more reliably observed for voices.

Method

Participants

128 participants aged between 18 and 35 years were recruited for online testing via Prolific.co. All participants were native speakers of English, had normal or corrected to normal vision and did not report having any hearing impairments. Racial identity was not controlled for. Ethical approval was given by the UCL Research Ethics Committee (Project ID number: SHaPS-2019-CM-030). One data set was unusable due to a technical error. Two further participants were excluded since they reported having some knowledge of Bulgarian. Five participants were excluded because catch trials indicated that they were not paying sufficient attention to the task (errors on > 20% of the catch trials). Of the remaining sample of 120 participants (mean = 26.9 years, *SD* = 6.1 years; 53 female), 115 participants responded correctly to all catch trials, 7 participants responded correctly for 90% of the trials and 1 participant for 80% of the trials. A further participant was excluded as they gave over 95% of the stimuli the same rating. Participants were randomly assigned to rate the faces and voices of either the male or the female identities for one of the three traits (trustworthiness, dominance, or attractiveness), leading to a sample size of 20 participants for trustworthiness and dominance and 19 participants for attractiveness.

Materials

In order to collect ratings for the face and voice of the same identities, we used a total of 200 video clips in Experiment 3: 10 videos of each of 20 unfamiliar white identities (10 female). All identities were Bulgarian celebrities, including actors, musicians, and television personalities, aged between 21 and 49 years. The video clips were downloaded from different social video sharing platforms (e.g., YouTube) and captured a variety of interviews, advertisements, and video blogs showing the selected identities. The video clips were originally uploaded between October 2010 and February 2020 and the 10 videos for each identity were recorded at 10 different and separate instances. Thus, the video clips include recordings of the 20 identities across several years. We extracted short excerpts from each video (mean duration = 1.30 seconds, SD = 0.27 seconds) that contained a full meaningful utterance. From these videos, we then extracted the muted video and audio separately. Through this process, we were able to include the same identities in the face and voice stimuli while additionally sampling the face and the voice recordings from the same original instance. The short video clips were cropped to show the face of the identity only in order to avoid additional sources of variability such as clothing or body cues. In order to compare the variability in social ratings from faces and voices, participants were either presented with a muted video clip (visual information only) or with an audio recording extracted from the video clip (auditory information only). The muted video clips showed the face in different poses, from different angles and with different emotional expressions. Like the video clips, the audio recordings featured the celebrities speaking in a range of speaking environments, featuring different speaking styles, emotional expressions, and conversation partners. As all celebrities spoke in a language unfamiliar to the participants (Bulgarian), we were also able to mitigate the effects of linguistic content.

Procedure

This experiment was also hosted on the Gorilla Experimenter (www.gorilla.sc, Anwyl-Irvine et al., 2018). Participants rated 100 voice recordings and 100 muted videos from either the male or the female identities for one of the three traits. Thus each participant completed 200 trials. Voice and face recordings were presented in blocks, with the order of the blocks being counterbalanced across participants. The rating procedure was otherwise identical to what is reported for previous

experiments. Variability estimates were obtained in the same way described in Experiment 2, with no permutations being necessary either for this experiment (see Experiment 1).

Results and Discussion

Inter-rater reliability

Ratings for all three social traits across both modalities showed good inter-rater agreement with Cronbach's α > .72 except for trustworthiness ratings attributed to male voices (Cronbach's α = .68). Since participants rated stimuli for one trait and of one sex (male or female) only, we calculated an ICC, separately for each trait, modality and stimulus sex. Table 1 shows the results of these analyses. Overall, there was a significant rater agreement in all cases.

Table 1. Intraclass Correlation Coefficients across each Trait, Modality and Stimulus Sex Condition in Experiment 3.

Trait	Modality	Stimulus Sex	ICC	95% Confidence Intervals
Trustworthiness	Auditory	Male	.672	[.57, .76]
		Female	.577	[.45, .69]
	Visual	Male	.748	[.67, .81]
		Female	.718	[.63, .79]
Dominance	Auditory	Male	.783	[.72, .84]
		Female	.812	[.75, .86]
	Visual	Male	.804	[.74, .86]
		Female	.867	[.83, .90]
Attractiveness	Auditory	Male	.772	[.70, .83]
		Female	.612	[.50, .71]
	Visual	Male	.895	[.90, .92]
		Female	.824	[.76, .87]

Note: Two-Way Random model, absolute agreement, all *p*s < .001

Comparing the within- and between-person variability in social trait ratings

The within- and between-person variability in social ratings attributed to the muted videos and audio recordings was calculated the same way as described for Experiment 2. The mean within- and between-person variability is shown in Figure 5, separately for each modality as well as female and male identities.

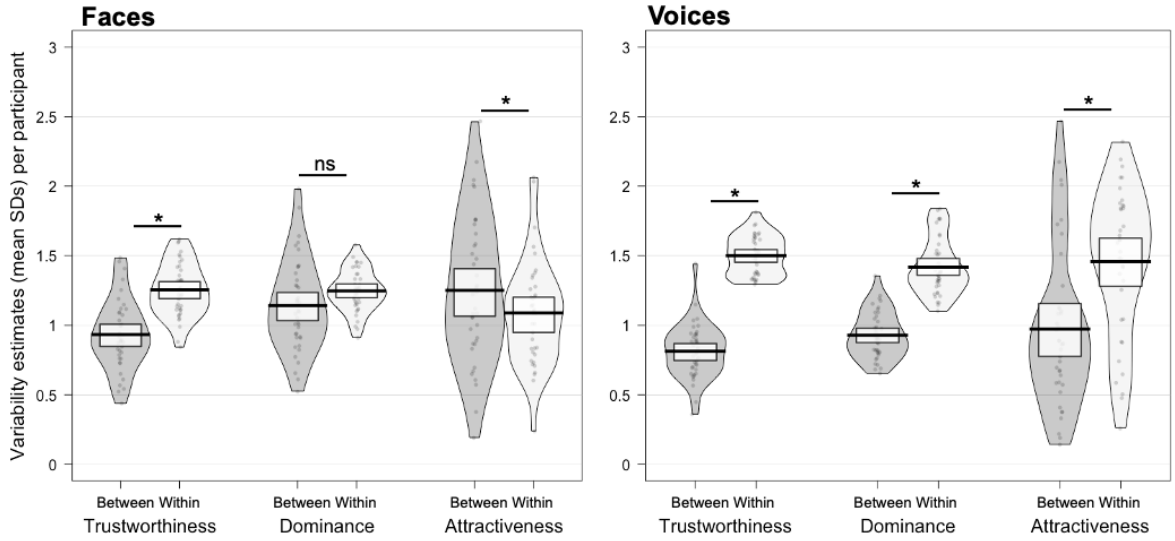


Figure 5. Within- and between-person variability in ratings of trustworthiness, dominance, and attractiveness across modality in Experiment 3. Boxes indicate the 95% confidence intervals. * indicates $p < .05$.

The LMMs had variability type (within vs between), social trait (trustworthiness, dominance, and attractiveness), modality (faces vs voices) and stimulus sex as fixed factors and participant as a random effect. Interactions between variability type, social trait and modality were modelled. Sex was not included in the interactions as it was not a factor of interest.

*lmer(variability estimate ~ variability type * trait * modality + sex + (1|participant)*

Significance of the main effects and interactions for the full model was established via log-likelihood tests in the same manner as for the models reported in the previous experiments. For plots of mean ratings per image and identity, please refer to Supplementary Figure 3.

The three-way interaction between social trait, modality and variability type was significant (χ^2 [2] = 7.20, p = .027). This three-way interaction appears to be driven by attractiveness ratings, where the between-person variability is larger for faces but not for voices (see also Experiment 1).

To test whether the two-way interactions between variability type and modality were significant for each social trait, we used the testInteractions function from the *phia* package (De Rosario-Martinez, 2015) in *R.* This analysis showed that this was indeed the case for each social trait (χ^2s[1] > 19.39*, p*s < .001). To then break down the two-way interaction, post-hoc tests were conducted on the model including the three-way interaction using *emmeans*. These analyses showed that within-person variability exceeded the between-person variability for faces for trustworthiness (t[348] = 5.48*, p* < .001) but not for dominance (t [348] = 1.79*, p* = .074). for Between-person variability exceeded within-person variability for attractiveness (t[348] = 2.73*, p* = .007). For voices, within-person variability exceeded the between-person variability for all three social traits (ts[438] > 8.33*, p*s < .001). Notably, the difference was more pronounced for voices than for faces, driving the two-way interactions for each social trait.

Thus, despite the considerable changes in stimulus materials, Experiment 3 again closely replicates the patterns of findings reported in Experiments 1 and 2 for both faces and voices. In this experiment, we were thus able to further extend our findings from the previous experiments to dynamic visual stimuli while minimising the effects of (intelligible) linguistic content on social trait judgements. We were furthermore able to replicate our finding from Experiment 2 and thus formally confirm that

within-person variability in social trait evaluations more consistently exceeds between-person variability for voices than for faces.

General Discussion

In a series of experiments, we examined the within- and between-person variability in social evaluations attributed to faces and voices along the three key dimensions of trustworthiness, dominance, and attractiveness taken from current models of face and voice perception (McAleer et al., 2014; Oosterhof & Todorov, 2008; Sutherland et al., 2013). Despite significant changes in the properties of our stimuli and some changes to the design (assignment of within- vs between-subject factors) across the three experiments, we find remarkably consistent patterns of results: As predicted, within-person variability in trait evaluations was either on par with, or exceeded, between-person variability in almost all cases for both faces and voices.

The results of our study therefore underline that the social trait impressions we form from glimpses of faces and snippets of voices do not directly reflect a truth about stable personality characteristics of a person (Todorov et al., 2015). Therefore, by measuring the degree of within-person variability and comparing it to the between-person variability in social ratings attributed to different instances of the same person, we introduce a new line of evidence directly exposing limitations of the 'kernel of truth' hypothesis (e.g., Berry, 1991; Penton-Voak, Pound, Little, & Perrett, 2006; Todorov et al., 2015) with regards to both face and voice impressions. More importantly, our approach is perhaps a more objective one since it does not rely on self-reported measures of personality frequently used to establish a 'ground truth'.

Instead of being cues to stable personality characteristics, social trait ratings in our experiments appear to be driven by less stable, situation-based appraisals of a person's behaviour (e.g., a person can be much more dominant in some situations compared to others), which are then misinterpreted as cues to an underlying stable personality. This misinterpretation forms a parallel with the 'fundamental attribution error' of interpreting what are actually situationally-driven behaviours as if they were personality dispositions, as noted in many studies in social psychology (Todorov, 2017). Despite the fact that trait impressions are unlikely to reveal real personality characteristics, we find good inter-rater agreement across all experiments implying that trait impressions are largely shared. This agreement was somewhat higher for faces than for voices, especially when estimated with a more sensitive measure such as ICC. A lack of perfect agreement therefore suggests that although substantially shared across different observers, trait impressions can still reflect a degree of personal taste (Hönekopp, 2006; Kramer et al., 2018) and might be guided by factors related to both stimulus and perceiver characteristics (Hehman et al., 2017).

For faces, we closely replicate findings from previous studies of variability in trait ratings across different images of the same person (Jenkins et al., 2011; Mileva et al, 2019; Sutherland et al., 2017; Todorov & Porter, 2014). We furthermore extend these findings to dynamic videos of faces, demonstrating certain similarities between the patterns of variability for trait evaluations formed based on 2D static and 3D dynamic person presentation. For voices, our findings are novel as there is to date no research quantifying the within-person variability in social trait perception, although they do align with the findings of substantial within-person variability in acoustic measures (Atkinson, 1976; Kreiman et al., 2015). We note, however, in light of findings from the identity recognition literature with such naturally-varying images and voice recordings, it is likely that participants were sometimes not aware of having rated multiple images of the same identities (Jenkins et al., 2011; Lavan et al., 2019). In everyday interactions, we are rarely presented with isolated images or voice recordings of unfamiliar people as was the case in this experiment. We are thus usually able to tie together different instances of a person's face or voice via multimodal and/or contextual cues. In

such real life settings, in which participants are thus able to evaluate different instances of the same unfamiliar identity as being the same person, we would predict generally lower within-person variability in social ratings than were observed in the current study. Our results nonetheless demonstrate that two different images of the same person - presented without contextual cues or explicit recognition - can give rise to social evaluations that are just as different as those attributed to images of two different people.

Moreover, we show that within-person variability exceeds between-person variability to a greater extent for voices than for faces. Why might hearing the voice of an unfamiliar individual create relatively more variable trait evaluations compared to seeing that same person's face? An important clue lies in the fact that we did not observe substantial differences in the overall degree of within- or between-person variability across experiments - despite using different types of stimuli and despite explicitly manipulating between-person variability in Experiment 2. These changes in the properties of the stimuli used across the three experiments might have been expected to have wide-ranging effects on trait evaluations, as they should introduce or remove different types of information about the face or the voice. For example, for voices, when listening to someone speaking in a language we cannot understand, we lose not only access to the linguistic content of what is being said. We also lack the expertise and experience to interpret some of the fine-grained cues that would enable us to make sense of some of the characteristics of a person if they spoke in a familiar language (e.g. do they have a standard or a regional accent? Are they likely to be highly educated or not?). At the same time, listening to speech in regional accents (Experiment 2) or a foreign language (Experiment 3) can boost stereotype-based interpretations that may modulate social trait judgements (e.g. Bayard, Weatherall, Gallois & Pittam, 2001). Similarly, videos of faces provide additional information through changes in eye gaze direction, head movements, facial expressions and speech-related mouth movements (see Roberts et al., 2009 for attractiveness). In our experiments, however, none of these changes in stimulus properties affected the overall relationship of within- and between-person variability in social trait evaluations.

We therefore suggest that the relatively enhanced difference of within- and between-person variability of vocal impressions reflects the fact that an individual can generate (intentionally or unintentionally) a wider range of vocal than facial cues. Voices are inherently dynamic; they only exist when a person is using their voice and there is no such thing as 'resting' or 'static' voice. Instead, the degree of voicing, pitch, speech rate, and the quality and tone of a voice can change from syllable to syllable, and situation to situation, to convey both linguistic and non-linguistic information (Lavan et al., 2019). Thus, researchers have to date struggled to pinpoint truly invariant diagnostic features of voices. Of course, faces can also change substantially on a short timescale through changes under the sender's control that include expression, gaze direction and viewing angle. However, some of the largest changes that make images of faces highly variable reflect properties of the external world, such as lighting type and direction, or internal changes such as state of health; all of these are often outside a person's immediate control. Moreover, unlike voices, faces do not disappear unless a person hides their face or turns away, and parts of the face can remain relatively static, with the net result of relatively more potentially diagnostic features being persistently present in faces. These overall differences in variability in the physical properties of a person's voice compared to a person's face may then translate into increased within-person variability in the perception of social traits from voices, as found in our experiments. This is, however, speculative and our data cannot provide evidence for this point. The relationship between variability in the physical properties of faces and voices and variability in perceived social traits is indeed likely complex, such that not all physical differences will be equally salient for the perception of traits (see also the lack of effect of our stimulus selection manipulations in Experiment 2).

The relatively higher within-person than between-person variability of voices that we have demonstrated also has important implications for recognising identity. Although unfamiliar identity perception from naturally-varying stimuli is challenging for both unfamiliar faces and voices, identity perception is generally less accurate for voices than for faces (Barsics, 2014; Stevenage & Neil, 2014). Social evaluations are largely based on the physical properties of images and acoustic properties of voices which makes the variability in social ratings a potential proxy for the general variability between the images and voice recordings used throughout the studies. This is further strengthened by the similarities between the results of present studies (based on trait ratings) for voice stimuli and previous studies measuring within- and between-person variability based exclusively on the acoustic properties of voices (Atkinson, 1976; Kreiman et al., 2015). It is therefore possible that our findings of higher within- than between-person variability for voices offer a ready explanation for the reported difficulties in voice *identity* recognition. That is, higher variability in the acoustic properties present in voice recordings of the same person (than in images of the same person) can easily account for voice recognition being a more challenging and error-prone task - both in terms of familiar and unfamiliar recognition as well as in learning new voices (Barsics, 2014; Barsics & Brédart, 2012; Brédart, Barsics, & Hanley, 2009; Hanley & Damjanovic, 2009; Hanley, Smith, & Hadfield, 1998). In this sense, our relative reliance on faces compared to voices for recognising identity may reflect differences in physical properties and result in the functional demands of everyday life (Young et al., 2020).

From this perspective, whilst the high within-person variability of voices may well be useful in the context of conveying important changeable social signals, it will at the same time create difficulties in learning new identities from vocal cues alone. However, as well as these direct consequences of high within-person variability, we might also speculate that the relationship of the within-person variability to between-person variability in trait evaluations may itself be linked to the ability to learn a new identity. For example, in cases where between-person differences in social trait evaluations exceed within-person variability, then social evaluations could be partially helpful cues for identity recognition. If an identity is most often perceived as relatively trustworthy, for example, the perception of consistent trustworthiness could help to discriminate this identity from others who are more often perceived as relatively untrustworthy. In this way, variability between images of faces might become partially useful information for face learning, rather than the 'noise' that such variability is so often taken to represent; we note that the possibility that variability can promote face learning has also been suggested in other accounts (Bruce, 1994; Burton, 2013; Kramer, Young & Burton, 2018) and demonstrated empirically (Andrews et al., 2015). As within-person variability more often and clearly exceeds between-person variability in trait evaluations for voices, however, trait evaluations from voices would therefore form a less reliable or potentially even misleading cue to identity than would be the case for faces. This could then also contribute to making identity perception from voices a more challenging task than for faces.

The interpretation that trait evaluations for voices are more reliant on and thus also more vulnerable to the effects of within-person variability fits well with some of our recent work showing an effect of familiarity on social evaluations from faces, where different images of the same familiar person were rated more similarly to one another than different images of the same identity when unfamiliar (Mileva et al., 2019). For faces at least, knowledge of a familiar person can modulate the interpretation of changeable signals. For voices, however, there were no differences in the way voice recordings of familiar and unfamiliar identities were rated; that is, having increased access to identity cues through familiarity did not affect variability in trait evaluations derived from the voice. We interpret this as evidence that increased variability leads to cues to identity being perceived more independently from cues to social traits for voices than faces (Lavan, Mileva, & McGettigan, 2020).

Aside from general modality differences, we also note some consistencies in the relationship between these two sources of variability for each social trait. This is particularly true for face stimuli (static and dynamic), where we report significantly more within- than between-person variability for ratings of trustworthiness and dominance and significantly more between- than within- person variability for ratings of attractiveness. The only exception to this pattern is in Experiment 2, where we find no difference between within- than between-person variability for ratings of attractiveness and trustworthiness. However, the data still follows the same general pattern. These findings fit well with the existing literature (Todorov & Porter, 2014) and would imply that face attractiveness is more stable than voice attractiveness. It furthermore suggests that impressions of attractiveness (from the face) may be formed in a different way than impressions of trustworthiness and dominance.

It is also worth pointing out that even though faces and voices are the identity cues that research on social perception has been consistently focussing on, they are of course not the only sources of trait-relevant information. Other cues, such as body shapes, personal names or even olfactory signals, can also readily form the basis of social trait evaluations (Hu et al., 2018; Mehrabian, 2001; Sorokowska, Sorokowski, & Szmajke, 2012). How impressions from these signals may vary and, crucially, how their processing compares to and interacts with the processing of faces and voices remains to date largely unclear. Thus, future work that takes cues beyond faces and voices into consideration is currently much needed in order to better understand person perception from a multi-modal perspective and to more accurately represent our everyday social interactions.

A key theoretical question that we aimed to address in this paper concerns the relation between face and voice perception. Because voices can signal a number of comparable personal characteristics to faces (such as age, sex, identity, and emotion), and because cognitive models of face perception were developed at a relatively early point (Bruce & Young, 1986; Burton, Bruce & Johnston, 1990; Haxby, Hoffman & Gobbini, 2000), an influential later idea has been that the voice forms a kind of 'auditory face', with a parallel functional organisation (Belin, Fecteau & Bédard, 2004; Campanella & Belin, 2007; Yovel & Belin, 2013). This analogy can be useful in a number of ways, one of which is to draw attention to the critical differences in the implications of social signals that are changeable from moment to moment and those that are relatively invariant (Haxby et al., 2000). In terms of the present discussion, trait impressions and social attributions can be highly changeable during a social encounter, whereas identity is invariant. This has important implications because many social signals are inherently somewhat ambiguous. The existence of changeable signals thus places a premium on creating a perceptual mechanism that can combine all available sources of information to achieve a reasonably rapid and accurate overall interpretation (Young, 2018; Young et al., 2020). Consistent with this, it is already evident that there is considerable overlap between the information that can be gained from faces and voices when forming a first impression of an unfamiliar individual (McAleer et al., 2014; Oosterhof & Todorov, 2008; Sutherland et al., 2013), and that we integrate these different sources so readily that it is difficult for perceivers to attend selectively to what is being gleaned from the face or voice itself (Mileva et al., 2018; Rezlescu et al., 2015). The present findings take us a step further by showing that, whilst it may represent a useful starting point, the auditory face analogy does not offer the kind of detail needed to understand within-person and between-person variability in trait impressions: Although we found broad similarities in the relationships of within- and between-person variability for faces and voices, clear modality differences are apparent that a narrow interpretation the ‘auditory face’ analogy cannot explain.

In sum, our series of experiments further highlights the importance of within-person variability for understanding both face and voice perception. Although humans appear to treat impressions of the social traits of others as proxies for evaluations of stable personality characteristics (Chen et al. 2016; Ert et al., 2016; Klofstad, 2016; Mileva et al., 2020; Sussman et al., 2013; Wilson & Rule,

2015), we show that transient changes in the physical properties of faces and voices lead to concomitant changes in social trait evaluations. Thus, we consistently observed that within-person variability in trait evaluations is at least on par with the between-person variability. The degree of variability of trait evaluations within- and between identities did, however, differ for faces and for voices. This offers intriguing insights into how person perception from faces and voices may not at all times proceed along entirely comparable processing pathways, qualifying the idea that the voice is an 'auditory face' (Belin et al., 2004; Campanella & Belin, 2007; Yovel & Belin, 2013) by showing some limitations to this often useful analogy. Future work will be required to further explore how different perceptual cues and types of information are accessed and weighted for different modalities. This can then provide further insights into the potential differences in the cognitive mechanisms and processing stages underpinning person perception from faces and voices.

References

Andrews, S., Jenkins, R., Cursiter, H., & Burton, A. M. (2015). Telling faces together: Learning new faces through exposure to multiple instances. *Quarterly Journal of Experimental Psychology, 68*(10), 2041-2050. https://doi.org/10.1080/17470218.2014.1003949

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*(1), 388-407. https://doi.org/10.3758/s13428-019-01237-x

Atkinson, J. E. (1976). Inter- and intraspeaker variability in fundamental voice frequency. *The Journal of the Acoustical Society of America, 60*(2), 440-445. https://doi.org/10.1121/1.381101

Baker, R., & Hazan, V. (2011). DiapixUK: Task materials for the elicitation of multiple spontaneous speech dialogs. *Behavior Research Methods, 43*(3), 761-770. doi: 10.3758/s13428-011-0075-y

Bar, M., Neta, M., & Linz, H. (2006). Very first impressions. *Emotion, 6*(2), 269-278. https://doi.org/10.1037/1528-3542.6.2.269

Barsics, C. G. (2014). Person recognition is easier from faces than from voices. *Psychologica Belgica, 54*(3), 244-254. doi: 10.5334/pb.ap

Barsics, C., & Brédart, S. (2012). Recalling semantic information about newly learned faces and voices. *Memory, 20*(5), 527-534. https://doi.org/10.1080/09658211.2012.683012

Bates, D., Mächler, M., Bolker, B., Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1-48. doi: 10.18637/jss.v067.i01

Bayard, D., Weatherall, A., Gallois, C., & Pittam, J. (2001). Pax Americana? Accent attitudinal evaluations in New Zealand, Australia and America. *Journal of Sociolinguistics, 5*(1), 22-49. https://doi.org/10.1111/1467-9481.00136

Belin, P., Bestelmeyer, P. E., Latinus, M., & Watson, R. (2011). Understanding voice perception. *British Journal of Psychology, 102*(4), 711-725. https://doi.org/10.1111/j.2044-8295.2011.02041.x

Belin, P., Fecteau, D., & Bédard, C. (2004). Thinking the voice: Neural correlates of voice perception. *Trends in Cognitive Sciences, 8*(3), 129-135. https://doi.org/10.1016/j.tics.2004.01.008

Berry, D. S. (1991). Accuracy in social perception: Contributions of facial and vocal information. *Journal of Personality and Social Psychology, 61*(2), 298-307. https://doi.org/10.1037/0022-3514.61.2.298

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*(4), 433-436. https://doi.org/10.1163/156856897X00357.

Brédart, S., Barsics, C., & Hanley, R. (2009). Recalling semantic information about personally known faces and voices. *European Journal of Cognitive Psychology, 21*(7), 1013-1021. https://doi.org/10.1080/09541440802591821

Bruce, V. (1994). Stability from variation: The case of face recognition. *Quarterly Journal of Experimental Psychology, 47A*(1), 5-28. https://doi.org/10.1080/14640749408401141

Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology, 77*(3), 305-327. https://doi.org/10.1111/j.2044-8295.1986.tb02199.x

Burton, A. M. (2013). Why has research in face recognition progressed so slowly? The importance of variability. *Quarterly Journal of Experimental Psychology, 66*(8), 1467-1485. https://doi.org/10.1080/17470218.2013.800125

Burton, A. M., Bruce, V., & Johnston, R. A. (1990). Understanding face recognition with an interactive activation model. *British Journal of Psychology, 81*(3), 361-380. https://doi.org/10.1111/j.2044-8295.1990.tb02367.x

Campanella, S., & Belin, P. (2007) Integrating face and voice in person perception. *Trends in Cognitive Sciences, 11*(12), 535-543. https://doi.org/10.1016/j.tics.2007.10.001

Chen, D., Halberstam, Y., & Yu, A. C. (2016). Perceived masculinity predicts U.S. Supreme Court outcomes. *PLoS ONE, 11*, e0164324. http://dx.doi.org/10.1371/journal.pone.0164324

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*(1), 98 -104. https://doi.org/10.1037/0021-9010.78.1.98

De Rosario-Martinez, H. (2015). phia: Post-Hoc Interaction Analysis. R package, version 0.2-1. https://CRAN.R-project.org/package=phia

Ert, E., Fleischer, A., & Magen, N. (2016). Trust and reputation in the sharing economy: The role of personal photos in Airbnb. *Tourism Management, 55*, 62-73. https://doi.org/10.1016/j.tourman.2016.01.013

Fuertes, J. N., Gottdiener, W. H., Martin, H., Gilbert, T. C., & Giles, H. (2012). A meta- analysis of the effects of speakers' accents on interpersonal evaluations. *European Journal of Social Psychology, 42*(1), 120-133. https://doi.org/10.1002/ejsp.862

Giles, H., Baker, S., & Fielding, G. (1975). Communication length as a behavioral index of accent prejudice. *Linguistics, 13*(166), 73-82. https://doi.org/10.1515/ling.1975.13.166.73

Giles, H., Wilson, P., & Conway, A. (1981). Accent and lexical diversity as determinants of impression formation and perceived employment suitability. *Language Sciences, 3*(1), 91-103. https://doi.org/10.1016/S0388-0001(81)80015-0

Hancock, P. J., Bruce, V., & Burton, A. M. (2000). Recognition of unfamiliar faces. *Trends in Cognitive Sciences, 4*(9), 330-337. https://doi.org/10.1016/S1364-6613(00)01519-9

Hanley, J. R., & Damjanovic, L. (2009) It is more difficult to retrieve a familiar person's name and occupation from their voice than from their blurred face. *Memory, 17*(8), 830-839. https://doi.org/10.1080/09658210903264175

Hanley, J. R., Smith, S. T., & Hadfield, J. (1998). I recognise you but I can't place you: An investigation of familiar-only experiences during tests of voice and face recognition. *The Quarterly Journal of Experimental Psychology: Section A, 51*(1), 179-195. https://doi.org/10.1080/713755751

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences, 4*(6), 223-233. https://doi.org/10.1016/S1364-6613(00)01482-0

Hehman, E., Sutherland, C. A. M., Flake, J. K., & Slepian, M. L. (2017). The unique contributions of perceiver and target characteristics in person perception. *Journal of Personality and Social Psychology, 113*(4), 513-529. https://doi.org/10.1037/pspa0000090.

Hönekopp, J. (2006). Once more: Is beauty in the eye of the beholder? Relative contributions of private and shared taste to judgments of facial attractiveness. *Journal of Experimental Psychology: Human Perception and Performance, 32*, 199-209. http://dx.doi.org/10.1037/0096-1523.32.2.199

Hu, Y., Parde, C. J., Hill, M. Q., Mahmood, N., & O'Toole, A. J. (2018). First impressions of personality traits from body shapes. Psychological Science, 29(12), 1969-1983. https://doi.org/10.1177/0956797618799300

Jenkins, R., White, D., Van Montfort, X., & Burton, A. M. (2011). Variability in photos of the same face. *Cognition, 121*(3), 313-323. https://doi.org/10.1016/j.cognition.2011.08.001

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). Analyzing and improving the image quality of stylegan*. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8110-8119). arXiv:1912.04958

Klofstad, C. A. (2016). Candidate voice pitch influences election outcomes. *Political Psychology, 37*(5), 725-738. https://doi.org/10.1111/pops.12280

Klofstad, C. A., & Anderson, R. C. (2018). Voice pitch predicts electability, but does not signal leadership ability. *Evolution and Human Behavior, 39*(3), 349-354. https://doi.org/10.1016/j.evolhumbehav.2018.02.007

Kramer, R. S. S., Mileva, M., & Ritchie, K. L. (2018). Inter-rater agreement in trait judgements from faces. *PloS One, 13*(8), e0202655. https://doi.org/10.1371/journal. pone.0202655

Kramer, R.S.S., Young, A.W. and Burton, A.M. (2018). Understanding face familiarity. *Cognition, 172*, 46-58. https://doi.org/10.1016/j.cognition.2017.12.005

Kreiman, J., Park, S. J., Keating, P. A., & Alwan, A. (2015). The relationship between acoustic and perceived intraspeaker variability in voice quality. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Kreiman, J., & Sidtis, D. (2011). *Foundations of voice studies: An interdisciplinary approach to voice production and perception*. West Sussex, UK: Wiley-Blackwell. doi:10.1002/9781444395068

Lavan, N. (2020, September 25). Within- vs between-person variaiblity in trait judgements. Retrieved from osf.io/7nx6s

Lavan, N., Burston, L. F., & Garrido, L. (2019). How many voices did you hear? Natural variability disrupts identity perception from unfamiliar voices. *British Journal of Psychology, 110*(3), 576-593. https://doi.org/10.1111/bjop.12348

Lavan, N., Burton, A. M., Scott, S. K., & McGettigan, C. (2019). Flexible voices: Identity perception from variable vocal signals. *Psychonomic Bulletin & Review, 26*(1), 90-102. https://doi.org/10.3758/s13423-018-1497-7

Lavan, N., Mileva, M., & McGettigan, C. (in press). How does familiarity with a voice affect trait judgements? *British Journal of Psychology*. https://doi.org/10.1111/bjop.12454

Lavan, N., Scott, S. K., & McGettigan, C. (2016). Impaired generalization of speaker identity in the perception of familiar and unfamiliar voices. *Journal of Experimental Psychology: General, 145*(12), 1604-1614. https://doi.org/10.1037/xge0000223

McAleer, P., Todorov, A., & Belin, P. (2014). How do you say 'Hello'? Personality impressions from brief novel voices. *PloS One, 9*(3). doi: 10.1371/journal.pone.0090779

Mehrabian, A. (2001). Characteristics attributed to individuals on the basis of their first names. Genetic, Social, and General Psychology Monographs, 127(1), 59-88.

Mileva, M., Kramer, R. S., & Burton, A. M. (2019). Social evaluation of faces across gender and familiarity. *Perception, 48*(6), 471-486. https://doi.org/10.1177/0301006619848996

Mileva, M., Tompkinson, J., Watt, D., & Burton, A. M. (2018). Audiovisual integration in social evaluation. *Journal of Experimental Psychology: Human Perception and Performance, 44*(1), 128-138. https://doi.org/10.1037/xhp0000439

Mileva, M., Tompkinson, J., Watt, D., & Burton, A. M. (2020). The role of face and voice cues in predicting the outcome of student representative elections. *Personality and Social Psychology Bulletin, 46*(4), 617-625. https://doi.org/10.1177/0146167219867965

Mileva, M., Young, A. W., Kramer, R. S., & Burton, A. M. (2019). Understanding facial impressions between and within identities. *Cognition, 190*, 184-198. https://doi.org/10.1016/j.cognition.2019.04.027

Nolan, F., & Post, B. (2014). The IViE corpus. In: J. Durand, U. Gut, & G. Kristoffersen (Eds.), *The Oxford handbook of corpus phonology* (pp. 475-485). doi: 10.1093/oxfordhb/9780199571932.013.025

Ohala, J. J. (1982). The voice of dominance. *The Journal of the Acoustical Society of America, 72*, S66. http://dx.doi.org/10.1121/1.2020007

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences, 105*(32), 11087-11092. http://dx.doi.org/10.1073/pnas.0805664105

Penton-Voak, I. S., Pound, N., Little, A. C., & Perrett, D. I. (2006). Personality judgments from natural and composite facial images: More evidence for a “kernel of truth” in social perception. *Social Cognition, 24*(5), 607-640. https://doi.org/10.1521/soco.2006.24.5.607

Read, D., & Craik, F. I. (1995). Earwitness identification: Some influences on voice recognition. *Journal of Experimental Psychology: Applied, 1*(1), 6-18. https://doi.org/10.1037/1076-898X.1.1.6

Rezlescu, C., Penton, T., Walsh, V., Tsujimura, H., Scott, S. K., & Banissy, M. J. (2015) Dominant voices and attractive faces: The contribution of visual and auditory information to integrated person impressions. *Journal of Nonverbal Behavior, 39*(4), 355-370. https://doi.org/10.1007/s10919-015-0214-8

Roberts, S. C., Saxton, T. K., Murray, A. K., Burriss, R. P., Rowland, H. M., & Little, A. C. (2009). Static and dynamic facial images cue similar attractiveness judgements. *Ethology, 115*(6), 588-595. https://doi.org/10.1111/j.1439-0310.2009.01640.x

Said, C. P., Sebe, N., & Todorov, A. (2009). Structural resemblance to emotional expressions predicts evaluation of emotionally neutral faces. *Emotion, 9*(2), 260-264. https://doi.org/10.1037/a0014681

Sorokowska, A., Sorokowski, P., & Szmajke, A. (2012). Does personality smell? Accuracy of personality assessments based on body odour. European Journal of Personality, 26(5), 496-503. https://doi.org/10.1002/per.848

Stevenage, S. V., & Neil, G. J. (2014). Hearing faces and seeing voices: The integration and interaction of face and voice processing. *Psychologica Belgica, 54*(3), 266-281. doi: http://dx.doi.org/10.5334/pb.ar

Stevenage, S. V., Symons, A. E., Fletcher, A., & Coen, C. (2020). Sorting through the impact of familiarity when processing vocal identity: Results from a voice sorting task. *Quarterly Journal of Experimental Psychology, 73*(4), 519-536. doi: 10.1177/1747021819888064

Sussman, A. B., Petkova, K., & Todorov, A. (2013). Competence ratings in US predict presidential election outcomes in Bulgaria. *Journal of Experimental Social Psychology, 49*(4), 771-775. https://doi.org/10.1016/j.jesp.2013.02.003

Sutherland, C. A. M., Oldmeadow, J. A., Santos, I. M., Towler, J., Burt, D. M., & Young, A. W. (2013). Social inferences from faces: Ambient images generate a three-dimensional model. *Cognition, 127*(1), 105-118. https://doi.org/10.1016/j.cognition.2012. 12.001

Sutherland, C. A. M., Young, A. W., Mootz, C. A., & Oldmeadow, J. A. (2015). Face gender and stereotypicality influence facial trait evaluation: Counter-stereotypical female faces are negatively evaluated. *British Journal of Psychology, 106*(2), 186-208. https://doi.org/10.1111/bjop.12085

Sutherland, C. A. M., Young, A. W., & Rhodes, G. (2017). Facial first impressions from another angle: How social judgements are influenced by changeable and invariant facial properties. *British Journal of Psychology, 108*(2), 397-415. https://doi.org/10.1111/bjop.12206

Tigue, C. C., Borak, D. J., O'Connor, J. J., Schandl, C., & Feinberg, D. R. (2012). Voice pitch influences voting behavior. *Evolution and Human Behavior, 33*(3), 210-216. https://doi.org/10.1016/j.evolhumbehav.2011.09.004

Todorov, A. (2017). *Face value: The irresistible influence of first impressions*. Princeton, NJ: Princeton University Press.

Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science, 308*(5728), 1623-1626. https://doi.org/10.1126/science.1110589

Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social attributions from faces: Determinants, consequences, accuracy, and functional significance. *Annual Review of Psychology, 66*, 519-545. https://doi.org/10.1146/annurev-psych-113011-143831

Todorov, A., & Porter, J. M. (2014). Misleading first impressions: Different for different facial images of the same person. *Psychological Science, 25*(7), 1404-1417. https://doi.org/10.1177/0956797614532474

Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological Science, 17*(7), 592-598. https://doi.org/10.1111/j.1467-9280.2006.01750.x

Wilson, J. P., & Rule, N. O. (2015). Facial trustworthiness predicts extreme criminal-sentencing outcomes. *Psychological Science, 26*(8), 1325-1331. https://doi.org/10.1177/0956797615590992

Woods, K. J., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics, 79*(7), 2064-2072. https://doi.org/10.3758/s13414-017-1361-2

Young, A. W. (2018). Faces, people and the brain: The 45th Sir Frederic Bartlett Lecture. *Quarterly Journal of Experimental Psychology, 71*(3), 569-594. https://doi.org/10.1177/1747021817740275

Young, A. W., & Burton, A. M. (2018). Are we face experts? *Trends in Cognitive Sciences, 22*(2), 100-110. https://doi.org/10.1016/j.tics.2017.11.007

Young, A. W., Frühholz, S., & Schweinberger, S. R. (2020). Face and voice perception: Understanding commonalities and differences. *Trends in Cognitive Sciences, 24*(5), 398-410. https://doi.org/10.1016/j.tics.2020.02.001

Yovel, G., & Belin, P. (2013). A unified coding strategy for processing faces and voices. *Trends in Cognitive Sciences, 17*(6), 263-271. https://doi.org/10.1016/j.tics.2013.04.004

Zebrowitz, L. A., & Montepare, J. M. (2008). Social psychological face perception: Why appearance matters. *Social and Personality Psychology Compass, 2*, 1497-1517. http://dx.doi.org/10.1111/j.1751-9004.2008 .00109.x

Supplementary Analysis 1: Experiment 1 - The effects of linguistic content on trait ratings

To assess whether the linguistic content of the words used in Experiment 1B systematically affected social trait evaluations (over and above the quality of the voice), we compiled arousal, valence, and dominance ratings for the words used in the study from a published database (Warriner, Kuperman & Brysbaert, 2013). Out of the 400 stimuli used in our experiment, no ratings were available for 30 of these stimuli. These stimuli were omitted from the analyses below.

To assess the effect of linguistic content on social trait evaluations, we computed the mean rating for each stimulus in our study, separately for each social trait (attractiveness, dominance, trustworthiness). We then ran three multiple linear regression analyses, one for each social trait, with the mean trait ratings per stimulus as the dependent variable and arousal, valence, and dominance ratings of the linguistic content of the stimuli as predictors.

For attractiveness ratings, the linguistic content of the words explained 3% of the variance in the trait ratings (p = .003). Dominance ratings of the linguistic content was a significant predictor for attractiveness ratings of the voices (β = .18, p = .004), arousal and valence

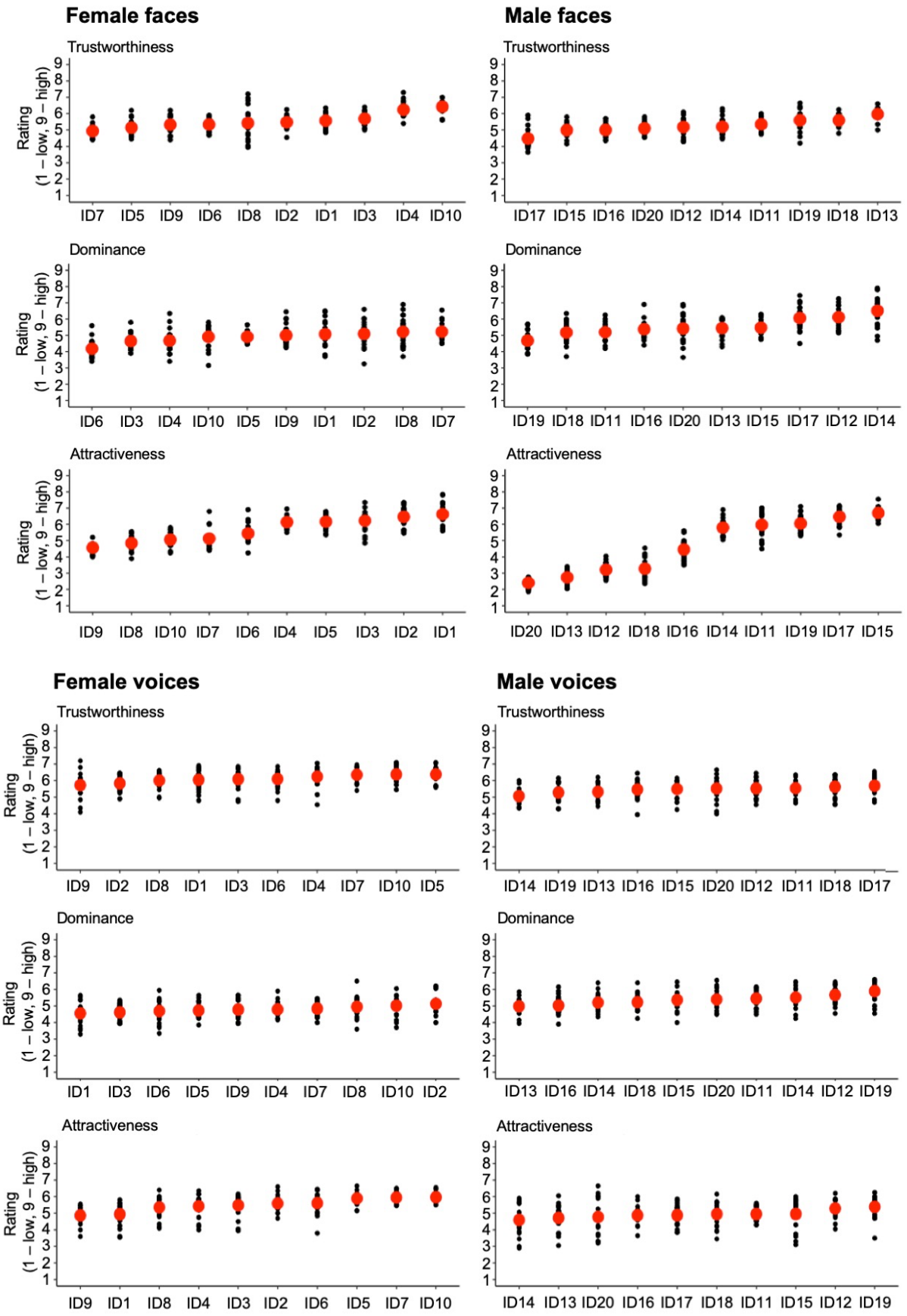
ratings of the linguistic content did not significantly predict attractiveness ratings (β < .03, p > .725).

For dominance ratings, the linguistic content of the words explained < 1% of the variance in the trait ratings (p = .473). None of the ratings of the linguistic content (not even dominance ratings) significantly predicted dominance ratings of the voices (β < .1, p > .128).

For trustworthiness ratings, the linguistic content of the words explained 8% of the variance in the trait ratings (p < .001). Dominance ratings of the linguistic content was a significant predictor for trustworthiness ratings of the voices (β = .19, p = .001). Similarly, valence ratings of the linguistic content were a significant predictor for trustworthiness ratings, although the predictive relationship was negative, such that negatively valenced words predicted higher trustworthiness ratings of voices (β = -.07, p = .044).

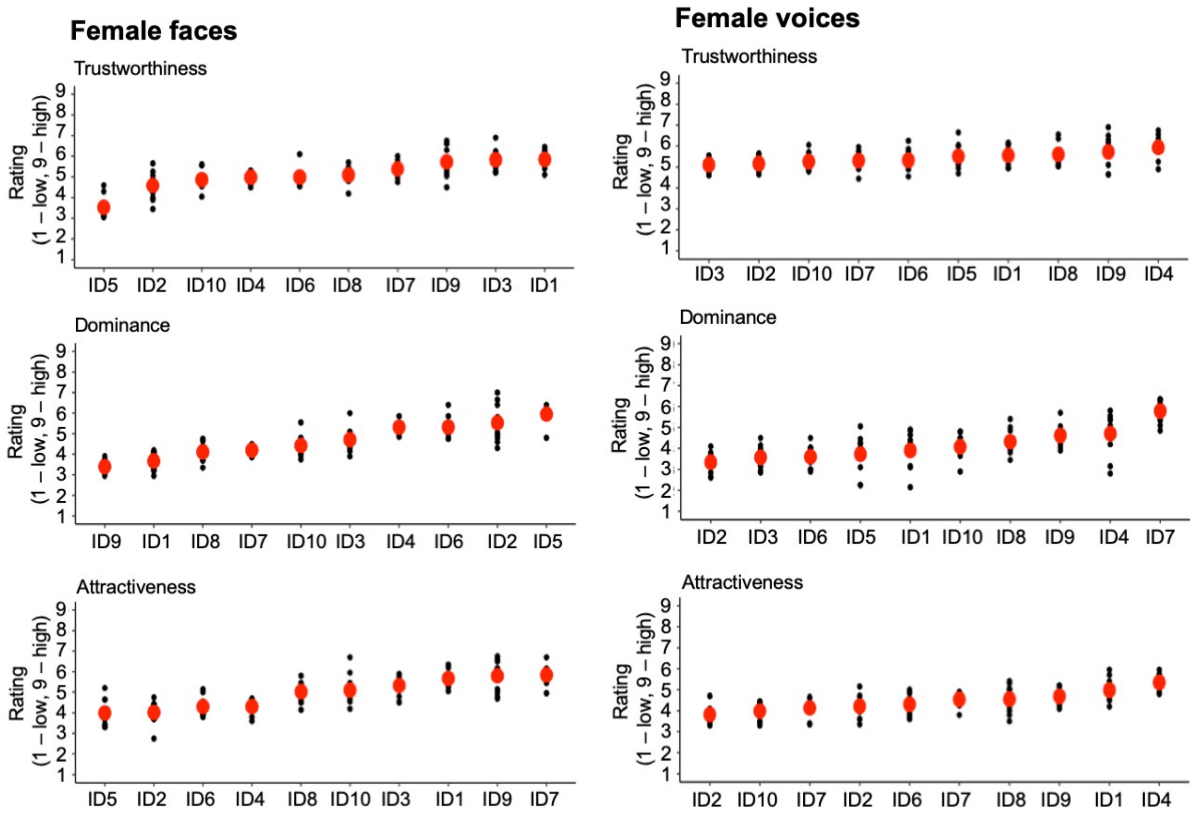
From these analyses, we conclude that linguistic content of the words can at times influence the trait evaluations of voices in our experiment. However, only a small portion of the variance is explained by the linguistic content (< 9%) and the pattern of predictive relationships appears at times counterintuitive (no relationship for dominance ratings; negative relationship of valence and trustworthiness ratings). We therefore conclude that the linguistic content is unlikely to have had a systematic, interpretable effect on the social evaluations collected in this experiment.

Supplementary Figure 1: Experiment 1 - Alternative plot of the within- and between-person variability



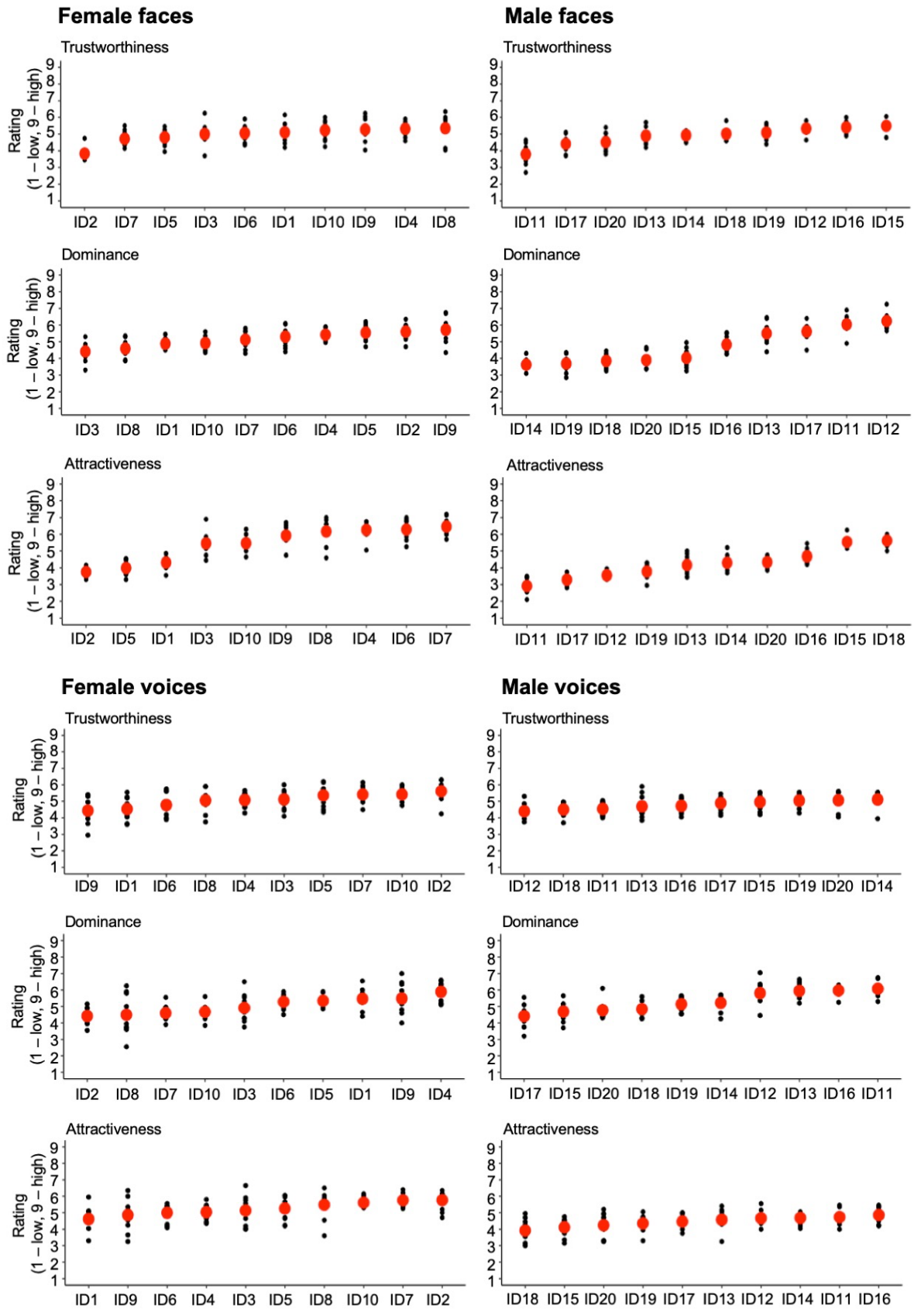
Supplementary Figure 2: Mean trait ratings plotted per item (black dots) and per identity (red dots). Higher within-person variability is indicated by a larger range in mean ratings per item. Higher between-person variability by a larger range of mean ratings per identity

Supplementary Figure 2: Experiment 2 - Alternative plot of the within- and between-person variability



Supplementary Figure 3: Mean trait ratings plotted per item (black dots) and per identity (red dots). Higher within-person variability is indicated by a larger range in mean ratings per item. Higher between-person variability by a larger range of mean ratings per identity.

Supplementary Figure 3: Experiment 3 - Alternative plot of the within- and between-person variability



Supplementary Figure 4: Mean trait ratings plotted per item (black dots) and per identity (red dots). Higher within-person variability is indicated by a larger range in mean ratings per item. Higher between-person variability by a larger range of mean ratings per identity.

References

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods, 45*(4), 1191-1207. https://doi.org/10.3758/s13428-012-0314-x